인체 골격의 정보의 기계학습을 통한 자세 인식 개선 방법

강민주, 류수경, 김나영, 이지은, 강제원*
이화여자대학교
*jewonk@ewha.ac.kr

요 약

본 논문에서는 개선된 자세 인식을 위한 학습을 통한 자세 인식 기법을 제안한다. 제안 자세 인식 기법은 영상의 모든 픽셀 값을 사용하지 않으며 인체의 골격의 위치 정보와 자세의 학습을 기반으로 한다. 최근 자세 인식기법에 다양한 기계 학습 기법을 적용하여 제스처 인식률을 높이는 연구가 진행되고 있지만 실시간 프레임에 적용하는데 한계가 있다. 반면 고차원의 특징점을 추출하여 신경망 학습방식을 이용하면 적은 계산량과 손쉬운 실행이 가능하다. 고차원의 특징점은 깊이 정보로부터 사람의 골격 정보를 이용해 추출하여 차원을 감소시키며 신경망 학습 방식에서는 각 자세에 대한 고차원의 특징점을 이용하여 자세의 학습을 진행한다. 신경망학습은 학습 단계에서는 미리 알려진 자세와 예측된 자세의 비교를 통해 오류를 최소화 하는 방향으로 학습을 진행하며, 판별 단계에서는 새로운 자세를 입력하여 고차원 특징점을 이용한 신경망 학습 기반의 제안 기술의 성능을 평가한다. 실험에 의하면 제안 기법은 약 96%의 자세 인식률을 보이고 자세 인식기법을 동작 인식으로 확장 가능성 또한 보인다.

1. 서론

자세(posture) 인식 기술은 동작(gesture) 인식 기술의 일부로 사용자와 신체의 움직임을 인식하여 컴퓨터와 상호 작용 하는 컴퓨터 비전 분야에서 중요한 연구 주제이다. 자세는 정지영상에서 취득한 몸 일부분의 움직임을 나타내며 동작은 시간적으로 구성된 여러 자세의 집합을 의미한다 [3]. 자세 및 동작인식 기술은 감시 시스템, 환자 모니터링 시스템의 지능형 시스템과 컴퓨터와 사람 사이의 실감 형 인터페이스에 널리 사용되고 있으며 사람의 다양한 동작을 보다 정밀한 인식률로 인지하기 위한 데 많은 연구가 진행되고 있다.

최근에는 기존의 컬러 영상과 다른 깊이 영상에 기반한 자세 및 동작 인식 기술들이 실용화되고 있다. 깊이 영상 센서는 3 차원 정보의 획득을 통하여 기존의 동작 인식의 어려움을 극복하는데 도움을 주고 있다. 대표적인 깊이 영상센서인 Microsoft 사의 Kinect 는 3 차원 정보로부터 사람을 인식 한 후 사람의 골격 정보(skeleton) 정보를 제공한다.

본 논문에서는 사람의 자세를 보다 효율적으로 인식하기 위하여 사람의 골격 정보로부터 특징을 추출하고 그를 이용한 신경망 학습 기반의 자세 인식 기법을 제안한다. 제안 기법에서는 Kinect 를 통해 영상의 깊이 정보에서 사람 의 관절 위치를 획득하고 이로부터 사람의 자세 및 동작을 학습하는 데 필요한 각도의 고차원 특징점을 추출한다. 추출한 각도 특징점은 신경망 네트워크 모델(neural network model)을 이용해 자세를 분류하게 된다. 제안 기법은 사람마다 가진 신체적 편차를 극복하여 높은 동작 인식 성능을 제공하며 자세 인식뿐만 아니라 동작인식에 대해서도 확장이 가능하다.

2. 기존기법

사람의 skeleton 정보를 사용한 자세 및 동작 인식에 관한 연구가 있다. Fujiyoshi at al. 은 사람의 동작을 분석하기 위해 움직이는 물체의 경계를 추출하여 star skeleton 를 생성 후 수직선과 팔, 수직선과 다리 사이의 각도로 부터 frame 에 따른 몸의 기울기 변화 추이를 계산하여 일부 동작에 대한 인식 연구를 수행하였다 [4]. 이 후 Kinect 를 이용하여 인체의 skeleton 정보를 파악하여 관절(joint)의 공간 좌표 값을 용이하게 획득할 수 있게 됨에 따라 3 차원 관절 정보를 이용한 동작 및 자세 인식에 관한 연구가 진행이 되었다. Raptis et al. 은 스켈레톤 정보를 이용한 춤 동작 분류 시스템을 제안하였다. 동작 인식을 위한 특징점으로 몸통 정보와 1 차 관절의 각도와 2 차 관절의 각도를 이용하였다 [5]. 그러나 한번의 판단만으로는 사람의 모션의 일치 여부를 판단하기에 정확성이 떨어지기 때문에 사람의 모션을 학습 한 후 보다 정확히 인지하는 학습을 통한 기법이 제시되었다. Cohen and Li 는 Support Vector Machine(SVM)을 이용하여 모션을 분류하는 기법을 제안하였으며[7] non-linear SV decision tree [1] 등을 통해 앉기, 걷기, 서있기 등의 동작을 구별하였다. 주성분 분석(Principal Component Analysis) 특징점 추출 방식을 사용한 신경망 학습 기반의 모션 인식 기법이 제안되었다. [2] Patsadu et al 은 자세 분류의 성능 분석을 위해 신경망, support vector machine, decision tree, naïve Bayes 학습 방식을 이용하였다. 신경망 학습 방식에서는 스켈레톤 20 개에 대한 3 차원 좌표를 그대로 특징점으로 사용하였으며 서있기, 앉아있기, 누워있기에 대한 동작 구별 성능이 뛰어났다[6]. 하지만 사람의 위치의 변화에 따라 좌표가 변하는 경우 정확한 결과를 내기 힘들며 다양한 골격을 가진 사람을 적용하였을 때 동일한 모션을 판별 하기 힘든 단점이 있었다.

3. 제안기법

제안하는 자세 인식 기법에서는 Microsoft 사의 Kinect 를 통해 획득한 인체의 skeleton 정보를 이용하여 사람 모션에서 특정 부분의 각도를 추출하여 특징점의 차원을 감소시키고 신경망 학습을 통해 사람의 편차에 강인한 신경망을 생성한다. 학습된 신경망에 새로운 자세가 포함되어 있는 영상은 학습한 신경망에 적용되어 어떤 자세인지 분류 된다. 그림 1 은 제안 기법의 순서도를 보여준다.

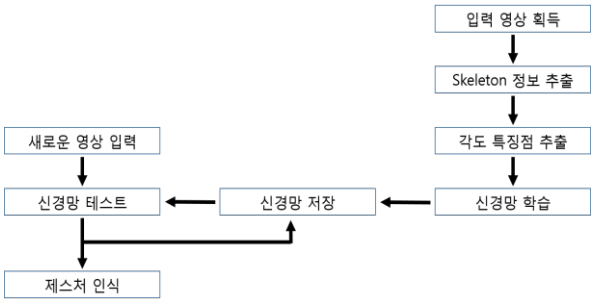

그림 1. 제안 기법의 순서도

Kinect 를 통해 위치 정보를 얻을 수 있는 사람의 관절은 그림 2 과 같다. 자세는 4 등분으로 나누며 각 부분마다 몸의 중심으로부터 3 개의 관절 정보를 추출한다. 그림 2 에서 P_j^i (i=1, 2, 3, 4, j=1, 2, 3) 는 i 번째 사분면의 j 번째 관절을 의미한다. 각 P_j^i 의 관절의 위치 정보는 (x_j^i, y_j^i, z_j^i) 의 3 차원 공간 좌표를 가진다. 제안 기법에서는 사람의 주요 모션을 4 등분하여 우측 상단(1 사분면), 좌측상단(2 사분면), 좌측 하단(3 사분면), 우측 하단(4 사분면)으로 나누어 각각의 관절 정보를 고려한다.

제안기법에서는 그림 3 과 같이 12 개의 고차원 특징점을 생성하여 각 자세를 구분 할 수 있는 요소로 사용한다. 즉 skeleton 정보 $P_j^i \in P$ 로부터 집합 P 내 원소들 간의 조합을 통해 특징점 집합 A 을 생성하여 자세 인식의 기반이 되는 특징점으로 사용하는데 이러한 과정을 특징점 추출이라고 한다. 제안 기법에서는 특징점 집합으로 각 관절 간 각도를 계산하여 원소로 한다.

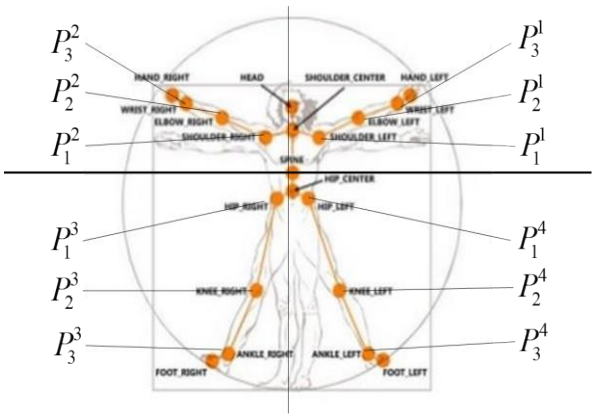

그림 2. 추출하는 skeleton 정보 그림

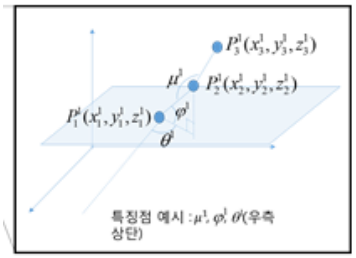

그림 3. 추출하는 각도 그림

자세 인식을 위한 특징점 추출을 위해 skeleton 정보의 world coordinate P_j^i 을 찾는다. 자세의 i 사분면에 위치한 우측 상단의 어깨, 팔꿈치, 손목의 skeleton 정보를 얻어 관절과 관절 사이의 각도 정보로부터 3 개의 특징점을 구한다. 그 첫 번째 특징점으로 그림 3 에서 보듯이 어깨-팔꿈치, 팔꿈치-손목 사이의 각도 u 를 어깨의 좌표가 P_1^i (x_1^i, y_1^i, z_1^i) 팔꿈치의 좌표가 P_2^i (x_2^i, y_2^i, z_2^i) 손목의 좌표가 P_3^i (x_3^i, y_3^i, z_3^i) 이라고 할 때, 다음과 같은 수식을 사용하여 각도 u^i 를 구한다.

$$u^i = 180^\circ - \{\arctan(\frac{y_3^i - y_2^i}{x_3^i - x_2^i}) - \arctan(\frac{y_1^i - y_2^i}{x_1^i - x_2^i})\} \times \frac{180^\circ}{\pi} \cdots (1)$$

두 번째 특징점은 팔 모션에서의 각도 φ 이며 skeleton 정보 중 어깨와 팔꿈치의 좌표를 사용한다. 어깨 P_1^i (x_1^i, y_1^i, z_1^i) 에서 $z = z_1^i$ 평면과 어깨-팔꿈치 좌표인 $p_1^i - p_2^i$ 직선이 이루는 각을 φ^i 라고 한다. φ^i 의 계산은 다음과 같다.

$$\varphi^i = \arctan(\frac{z_2^i - z_1^i}{x_2^i - x_1^i}) \times \frac{180^\circ}{\pi} \cdots (2)$$

세 번째 특징점은 θ 이며, 각도 스켈레톤 정보 중 어깨와 팔꿈치의 좌표를 사용하여 어깨 P_1^i (x_1^i, y_1^i, z_1^i) 에서 $y = y_1^i$ 평면과 어깨-팔꿈치 좌표인 $p_1^i - p_2^i$ 직선이 이루는 각을 θ^i 라고 한다. θ^i 의 계산은 다음과 같다.

$$\theta^i = \arctan(\frac{y_2^i - y_1^i}{x_2^i - x_1^i}) \times \frac{180^\circ}{\pi} \cdots (3)$$

이어서 각 4 분면의 특징 각도 각 3 개씩 동일한 방식으로 추출하여 집합 A 를 구성한다.

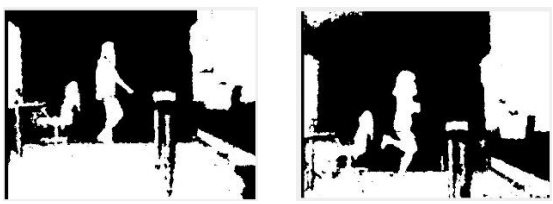

그림 5. 걷는동작(좌)와 달리는 동작(우)의 깊이 정보

제안 기법에서는 기계학습 중 신경망 학습을 통해 자세을

인식하는 방식을 보인다. 신경망 학습은 감독학습 중 하나로 입력 데이터에 X 에 대해 올바른 정답 Y 를 학습하는 방식으로 제안방식에서의 그림 4 에서의 입력 데이터 X 는 12 개의 각도 특징점(u^1,φ^1,θ^1,u^2, φ^2, θ^2,u^3, φ^3, θ^3,u^4,φ^4,θ^4)과 각 자세의 Label 이 정답 Y 가 된다. 즉, N 개의 자세에 각각 Label 을 붙여 올바른 정답인 Lable1~Label N 로 정의 할 수 있으며 Label 1~Label N 자세를 다양한 사람에게서 입력 받은 후 각 자세에 대한 12 개의 특징점(입력 X) 들을 입력 후 학습하면 Label1 ~ Label N 자세를 판별할 수 있는 그림 3 와 같은 신경망네트워크를 생성한다.

제안 기법에서의 움직임 판별의 목적은 현재 입력 자세가 학습한 Y 개의 자세 중 어떤 자세인지 인식하는 것이다. 입력 영상에서 학습에 사용한 동일한 절차로 특징점 X=12 개를

추출하고 학습을 이용해 취득한 파라메터 집합 W(신경망 네트워크의 경우 노드의 구조와 노드 간 연결 계수)을 이용하여 학습한 신경망 네트워크에 입력하면 최종 출력 노드에서 Label1 ~ Label N 자세 중 1 개로 판별 한다.

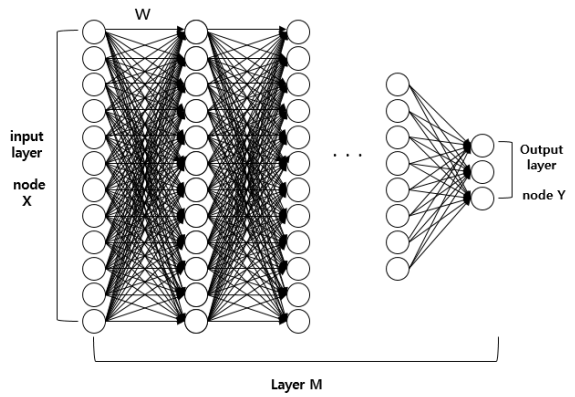

그림 4. 신경망 네트워크 예시

4. 실험결과

제안 기법의 효율적인 자세 인식 성능을 보이기 위하여 사람의 자세 인식하는 실험과 사람의 동작 또한 인식하는 실험을 구성하였다. 3 명의 사람으로 실험을 진행하였으며 3 개의 자세(옆으로 나란히, 소총, 가스)를 구분하는 실험과 2 개의 동작(걷기, 달리기)를 구분하는 실험을 진행하였다. 이미지의 크기는 480x640pixel 이며 이미지는 깊이 정보에서 skeleton 이 추출 될 때 마다 1.5fps 의 속도로 이미지를 캡쳐하였다. 자세 인식 실험에서는 총 163 개의 영상 샘플(옆으로 나란히 58 개, 소총 60 개, 가스 55 개)와 액션 인식 실험에서는 총 600 개의 영상 샘플 (걷는 동작 300 개, 달리는 동작 300 개)를 사용하였다. 자세 인식 실험에서는 학습에 사용된 데이터 121 개, validation 에 사용된 데이터 26 개, 테스트에 사용된 데이터 25 이고 신경망 은닉층의 개수는 10 개 혹은 15 개로 설정하였다. 액션 실험에서는 학습에 480 개의 데이터를 사용하였으며 validation 과 test 에는 각각 60 개 씩 사용하였으며 신경망 은닉층의 개수를 15 개로 설정하였다. 그림 5 는 걷는 동작과 달리는 동작의 깊이 정보의 예시를 보여준다.

표 1 은 제스처(옆으로 나란히, 소총, 가스 동작) 인식 실험의 결과를 표 2 는 액션(걷기, 달리기) 인식 실험 결과를 보여준다. 각 동작의 인식률은 아래의 계산 방식을 사용하였다.

$$\text{자세동작의 인식률}(\%) = \frac{\text{올바르게 인식 한 자세}}{\text{해당 자세의 전체 영상 수}}$$

표 1 에서 사용한 동작은 군대의 수신호로 사용되고 있는 옆으로 나란히, 소총, 가스 동작이었으며 타 동작과 달리 상대적으로 다리의 움직임은 큰 영향을 미치지 않고 상체의 움직임이 제스처의 주요 의미를 담고 있다. 가스 동작과 같이 한 손으로 코를 잡는 듯한 복잡한 동작 또한 여러 사람을 적용하였을 때에도 96.7%의 높은 인식률을 보였다. 표 2 에서 또한 걷거나 달리는 연속 동작을 구분해 내는 인식률이 약 90%에 가까움을 보여주고 있다. 이는 신경망 학습 방식 기반의 자세 인식 방법이 동작 인식 방법에도 확장 될 수 있음을 보여준다..

동작(제스처)	인식율
옆으로 나란히	98.2%
소총	94.8%
가스	96.7%

표1 자세 인식 성능 표

동작(액션)	인식율
걷기	90.7%
달리기	88.8%

표2 움직임 인식 성능 표

5. 결론

본 논문에서는 사람의 골격 구조를 추출하여 12 개의 각도를 특징점으로 삼아 자세 인식을 수행 할 수 있는 기법을 제안하였다. 제안 기법에 따르면 자세를 취하고 있는 사람을 각 자세의 주 표현 부분이라고 할 수 있는 오른쪽 팔, 왼쪽 팔, 오른쪽 다리, 왼쪽 다리를 기준으로 4 등분 한다. 각 부분의 양 팔과 다리 부분의 관절 12 개의 공간 좌표를 얻은 뒤 12 개의 각도를 계산하고, 고차원의 특징점을 사용하여 신경망 학습을 한 후 복잡한 동작 임에도 약 96%의 제스처 인식률을 보였다. 또한 약 90%의 동작 인식률의 높은 성능을 보이면서 제안 방식이 동작 인식 방법에도 확장 가능하다는 것을 알 수 있다.

6. 참고문헌

[1] Zhao, Haiyong, and Zhijing Liu. "Human action recognition based on non-linear SVM decision tree." Journal of Computational Information Systems 7.7 (2011):

2461-2468.

[2] Yu, Hui, et al. "Human motion recognition based on neural network."Communications, circuits and systems, 2005. Proceedings. 2005 international conference on. Vol. 2. IEEE, 2005.

[3]Aggarwal, Jake K., and Michael S. Ryoo. "Human activity analysis: A review."ACM Computing Surveys (CSUR) 43.3 (2011): 16.

[4]Fujiyoshi, H.; Lipton, A.J., "Real-time human motion analysis by image skeletonization," Applications of Computer Vision, 1998. WACV '98. Proceedings., Fourth IEEE Workshop on , vol., no., pp.15,21, 19-21 Oct 1998

[5] Michalis Raptis, Darko Kirovski, and Hugues Hoppe. 2011. " Real-time classification of dance gestures from skeleton animation. " In Proceedings of the 2011 ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA '11), Stephen N. Spencer (Ed.). ACM, New York, NY, USA, 147-156.

[6] Patsadu, O.; Nukoolkit, C.; Watanapa, B., "Human gesture recognition using Kinect camera," Computer Science and Software Engineering (JCSSE), 2012 International Joint Conference on , vol., no., pp.28,32, May 30 2012-June 1 2012

[7] Cohen, Isaac, and Hongxia Li. "Inference of human postures by classification of 3D human body shape." Analysis and Modeling of Faces and Gestures, 2003. AMFG 2003. IEEE International Workshop on. IEEE, 2003.